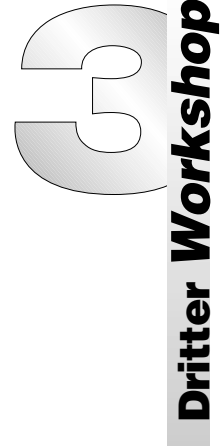


Das Internet und das WWW



Die Welt von Opera ist das Internet – ein boomendes Etwas, das wir hier in einem kleinen Exkurs beschreiben wollen. Insbesondere das so genannte World Wide Web oder WWW (das fälschlicherweise oft mit dem Internet gleichgesetzt wird) soll dabei im Mittelpunkt stehen. Das WWW mit seinen bunten Webseiten ist ja für die meisten Internetanwender der interessanteste Aspekt des Netzes der Netze (wie das Internet auch gerne genannt wird). Aber auch die anderen Dienste des Internets müssen beachtet werden, zumal Opera für deren Nutzung viele Schmankerl bereitstellt. Elektronische Post, Diskussionsforen, das Ausführen von Programmen auf entfernten Computern oder das Laden von Dateien und Programmen von – über die ganze Welt verteilten – Rechnern auf den eigenen Computer und wieder zurück. All das sind Dienste im Internet, die erst einmal jenseits des WWW anzusiedeln sind. Zumal sie fast alle viel älter als das WWW sind. Viel von der Struktur des Internets wird über dessen Geschichte deutlich.

3.1 Ein kurzer Abriss der Internet-Historie

Die Geschichte des Internets geht bis in die Zeit des Kalten Kriegs Anfang der 60er Jahre des letzten Jahrhunderts zurück. Die Invasion in der Schweinebucht in Kuba, die geplante Stationierung von Atomwaffen auf Kuba durch die damalige UdSSR, diverse Aktionen in Süd- und Mittelamerika durch unterschiedliche Geheimdienste – all das und die fast paranoide Furcht amerikanischer Militärs vor einem Atomschlag durch die UdSSR veranlasste in den späten 60er Jahren die Vereinigten Staaten von Amerika, eine Organisation namens Advanced Research Projects Agency (ARPA) zu beauftragen, für das US-Verteidigungsministerium (Department of Defence) ein Netzwerksystem mit über ein möglichst großes Gebiet verteilten Rechnern zu entwickeln, das selbst nach einem atomaren Erstschlag der UdSSR noch funktionieren und den Gegenschlag organisieren sollte. Die Entfernung zwischen den einzelnen militärischen Rechnersystemen sollte deshalb so groß wie möglich sein, damit sie nicht gleichzeitig außer Gefecht gesetzt werden konnten. Das erste greifbare Resultat war ein Forschungsnetzwerk aus vier miteinander verbundenen Großrechnersystemen, das im Jahre 1969 zusammengeschlossen wurde.

Bei der Konzeption des ARPANET (so der Name) war man von vornherein davon ausgegangen, dass inhomogene Rechnersysteme zusammenarbeiten mussten. Insbesondere bei der Verbindung sollten heterogene Wege möglich sein. Genau genommen kam der Verbindung zwischen diesen unabhängigen Rechnersystemen die zentrale Bedeutung zu. Diese Koppelung zwischen den einzelnen Systemen sollte unter allen Umständen funktionieren – sogar beim Ausfall eines beliebigen Rechners im Netz oder bei der Unterbrechung einer Anzahl von Leitungswegen. Das erforderte neben festen Standleitungen zur Verbindung der beteiligten Rechner die Möglichkeit der Koppelung über flexible Leitungen. Etwa über Telefonleitungen, die durch ihre Verflechtung über den gesamten Kontinent quasi unzerstörbar waren (und sind). Bei der Nutzung eines solchen Netzes muss beachtet werden, dass eine Verbindung zwischen zwei Teilnehmern einer Kommunikation in der Regel aus einzelnen Teilstrecken zusammengesetzt wird. Diese Zusammensetzung erfolgt an Knotenpunkten (den Ämtern). Da für eine maximale Fehlertoleranz und Sicherheit des Netzes immer mehrere Wege zum Austausch von Daten zwischen sämtlichen Computern im Netzwerk zur Verfügung stehen mussten, war ein Konzept erforderlich, in dem zu übertragende Informationen in der Lage sein mussten, quasi selbständig den Weg zum Adressaten zu finden. Genau genommen entwickelte man ein System, in dem Daten so gekennzeichnet werden, dass an Knotenpunkten in den Verbindungslinien die Daten so ausgewertet werden konnten, dass sie auf dem optimalen freien Weg zum nächsten Knotenpunkt geschickt werden konnten. Dies wiederum erforderte die Zerlegung von Informationen in kleinere Datenpakete, die unabhängig voneinander verschickt werden konnten. Auf dieser Idee basiert auch heute noch die Kernfunktionalität des Internets (gleich mehr dazu).

In der Anfangsphase wuchs das ARPANET langsam (was in Anbetracht der zu der Zeit immensen Kosten sicher verständlich ist) und bestand 1972 aus ungefähr 50 miteinander verbundenen Rechnersystemen. Insbesondere zwang zu dieser Zeit der Kostendruck die Militärs, das Netz für zivile Zwecke zu öffnen. Damit konnten lange Verbindungsstrecken zwischen militärischen Rechnern durch zivil finanzierte Teilstrecken ergänzt werden. Besonders die National Science Foundation (NSF), eine Dachorganisation verschiedener Bildungs- und Forschungseinrichtungen in den USA, zeigte schnell Interesse am Internet und sorgte dafür, dass über die Zeit zahlreiche Universitäten und andere Forschungseinrichtungen an das Internet angeschlossen wurden und die Verflechtung immer komplexer wurde. Mit allen Vor- und Nachteilen. Einmal wurde das Netz immer größer und damit immer weniger zerstörbar, andererseits erlangten damit immer mehr Personen Zugang zum Netz, die nicht unbedingt als militärisch zuverlässig zu betrachten waren. Schüler und Studenten hauptsächlich, die im Netz ihren Spiel- und Forschungstrieb auslebten. Der Film »War games« zeigt die Resultate solchen Spieltriebs – ohne es zu ahnen, startet ein Schüler die Simulation eines Angriffs mit Atomraketen und setzt damit fast den dritten Weltkrieg in Gang (der Film beruht auf wahren Begebenheiten). Die Militärs begannen auf jeden Fall, um die Sicherheit ihrer Geheimnisse und Systeme zu bangen, weshalb Anfang der 80er Jahre große Teile des militärischen Bereichs in ein eigenes Netz, das

MILNET, ausgliedert wurden. Der zivile Teil wurde Internet genannt. Dabei ist die Frage, woher der Name »Internet« kommt, nicht ganz unumstritten. Es gibt mehrere Antworten. Folgende ist aber die wahrscheinlichste: Internet steht für »Interconnected Networks«, was übersetzt »Netz der verbundenen Netze« bedeutet.

Kommen wir zurück zu der Verbindung von Rechnern über flexible Wege und Leitungen. Um dies überhaupt bewerkstelligen zu können, wurden verschiedene neue Techniken im Bereich der Hardware, vor allem aber Protokolle entwickelt, über die sich die Rechner verständigen konnten und die unabhängig von der eigentlichen Rechnerarchitektur und dem normalen Befehlssatz des jeweiligen Betriebssystems waren. Ein Protokoll ist dabei einfach als eine Vereinbarung zu verstehen, wie bestimmte Prozesse abzulaufen haben, und die alle an dem Prozess beteiligten Partner kennen und befolgen. Viele im Internet gebräuchliche Protokolle basieren auf Unix-Protokollen. Der Grund dafür ist, dass das ursprünglich im ARPANET hauptsächlich verwendete Betriebssystem Unix war, und so waren die am Anfang verwendeten Protokolle Unix-Datenübertragungsprotokolle. Nach diversen Experimenten wurde 1973 mit IP (Internet Protocol) ein Transportprotokoll präsentiert, das den plattformunabhängigen Transport von Daten über beliebige Wege und Leitungen garantieren sollte. Dazu wurden die Daten im Wesentlichen in kleine Pakete zerlegt und jedes dieser Pakete mit einer Adresse versehen, zu welchen Rechner es weitergeleitet werden soll (das impliziert, dass jeder Rechner im Netz eine eindeutige Adresse hat). Dazu kam als eine der weiteren Informationen die Adresse des Absenders, damit eine Antwort zurückgeschickt werden konnte.

Das Protokoll zeigte gute Ansätze, aber auch noch einige Schwächen, weshalb bereits Anfang 1974 eine darauf aufsetzende Verbesserung eingeführt wurde – das Erweiterungsprotokoll TCP (Transmission Control Protocol), das eine weitestgehend fehlerfreie Übertragung gewährleistete.

TCP funktioniert nur mit IP als Unterbau, gibt aber diesem erst seine Übertragungssicherheit. Und so kommt es, dass heute beide meist nur noch in Verbindung verwendet werden (TCP/IP). Seit 1983 ist TCP/IP das Internet-Standard-Protokoll für Direktverbindungen. TCP/IP ist ein sehr leistungsfähiges Transportprotokoll, aber eben nur ein reines Transportprotokoll. Man kann sich die Situation ganz gut vorstellen, wenn man sie mit dem Postsystem vergleicht. TCP/IP ist in dieser Vorstellung das System aus Briefkästen, Abholern, Verteilern und Zustellern, das die Übermittlung von Briefen ermöglicht. Was dann aber beim Empfänger mit dem Inhalt der Briefe angefangen werden soll, das steht auf einem anderen Blatt. Stellen Sie sich einfach einmal vor, was Sie mit einem Brief anfangen, der in Chinesisch geschrieben ist. Er wird Ihnen zwar zugestellt (wenn er korrekt und lesbar adressiert wurde), aber Sie werden wahrscheinlich die Information innerhalb des Umschlags nicht nutzen können.¹ Auch hier müssen wieder Vereinbarungen getroffen werden, welche Informationen im Brief selbst was zu bedeuten haben. Das bedeutet

¹ Es sei denn, Sie beherrschen Chinesisch.

erneut die Vereinbarung von Protokollen, die als Dienstprotokolle oder Internetdienste bezeichnet werden.

Die Dienstprotokolle, die auf dem TCP/IP-Übertragungssystem aufsetzen, sind Dienste wie Telnet, FTP, E-Mail oder auch das WWW. Insbesondere der letztgenannte Dienst – das WWW – hat dem Internet seine heutige Popularität verschafft. Das im Wesentlichen aus HTML (HyperText Markup Language – mehr dazu siehe etwas weiter unten) aufgebaute World Wide Web gibt es seit 1991. Die ersten Ansätze gehen allerdings bereits bis in den März 1989 zurück. Der Grund für das Entstehen des WWW war der, dass Ende der achtziger Jahre viele Internetanwender – meist Wissenschaftler aus EDV-fremden Sparten – nicht in der Lage oder gewillt waren, sich mit der relativ komplizierten Befehlsstruktur der bis dahin vorhandenen Internetdienste auseinander zu setzen. Verlangt wurde ein System, das auch Nicht-EDV-Freaks die Möglichkeit eröffnen sollte, Daten und Informationen über das Internet auszutauschen und in weltweiten Datenbeständen zu recherchieren. Resultat dieser Forderungen war ein Konzept, in dem im Internet verteilte Inhalte über ein vernetztes Hypertext-System in einer grafischen und vor allem einfach zu bedienenden Oberfläche (also mit der Maus anklickbare Texte, die als aktive Querverweise zu anderen Inhalten führen) verbunden werden. Für dieses Hypertext-System musste vor allem ein neues Protokoll entwickelt werden – HTTP (HyperText Transfer Protocol). Eigens für die Umsetzung des Projekts wurde eine Organisation namens W3C (<http://www.w3.org>) gegründet.

Mit der ersten Version von HTML präsentierte das W3C 1991 einen ersten einfachen Standard, in dem zunächst nur die Möglichkeit bestand, relativ simple Hypertext-Systeme im Internet aufzubauen. Diese konnten jedoch bereits wie geplant neben einfachem Text und Hyperlinks Grafiken enthalten. Insbesondere war damit das daraus entstehende WWW den bisherigen Internetdiensten in Bezug auf Anwenderfreundlichkeit überlegen.

Richtig populär wurde das WWW 1993 durch den kostenlos verteilten Mosaic-Browser, den ersten WWW-Browser mit einer grafischen Oberfläche. Lange Zeit stand der Mosaic-Browser als Synonym für die Darstellungssoftware des WWW. Insbesondere ist er der direkte Vorfahr des Netscape Navigators, den der Entwickler des Mosaic-Browsers – Marc Andreessen – 1994 entwickelt hatte, nachdem er NSCA verlassen und mit Jim Clark die Firma Netscape gegründet hatte. Dieser Browser hat im WWW – wie schon angedeutet – das letzte Jahrzehnt des vergangenen Jahrhunderts dominiert.

3.2 Wer steht hinter dem Internet?

Obwohl das Internet ursprünglich als militärisches System konzipiert wurde, gibt es schon lange mehr keine oberste Instanz, kein Kontrollorgan, keine Organisation zur Zensur. Zwar gehören einzelne Teile des Internets (Hosts, WANs, LANs, der private Rechner, Teile der Leitungen etc.) irgendwelchen Besitzern, das Netz als Ganzes gehört jedoch nie-

mandem. Es gibt auch keine zentrale Gesellschaft, die das Internet überwacht und lenkt. Das ist auch technisch unmöglich. Das Netz war ja genau so konzipiert worden, dass es weder zerstörbar noch ausspionierbar oder kontrollierbar ist. Hier haben die Väter des Internetkonzepts ganze Arbeit geleistet. Das Internet ist im besten Sinne ein chaotisches System. Es existieren jedoch einige Gruppen und Organisationen, die sich mit Entwicklungen im Internet befassen und versuchen, Standards zu definieren und Absprachen zur Kontrolle zu treffen und das Internet (in gewissen Grenzen) lenken. Dazu zählen etwa NIC (Internet Network Information Center – <http://www.internic.net>) als eine internationale Organisation zur Verwaltung von Domain-Namen und Internetadressen (siehe etwas weiter unten) oder das bereits erwähnte W3C, das sich um die Standards im WWW (etwa HTML, aber auch Style Sheets, JavaScript oder XML) kümmert. Diese Organisationen sind allerdings auf die freiwillige Kooperation der Internetteilnehmer (vor allem der einflussreichen kommerziellen Unternehmen) angewiesen. Und dass so eine freiwillige Kooperation nicht oft funktioniert, zeigen insbesondere die Standards, die das W3C herausgibt. Die großen Browser-Hersteller kümmern sich nicht viel um die Empfehlungen des W3C (mit Ausnahme von Opera!).

3.2.1 Die Adressierung im Internet

Wie schon angedeutet, basiert die Informationsübermittlung im Internet auf eindeutigen Adressen von Rechnern und der Zerlegung von Informationen in einzelne Datenpakete, die mit diesen Adressen versehen werden. Das TCP/IP-Protokoll bietet dafür die Basis. Diese so zerlegten Informationspakete müssen auf ihrem Weg vom Sender zum Empfänger oft mehrere Knoten passieren. Für diese Weiterleitung der Daten gibt es im Internet spezielle Server mit Adressinformationen und einer Weiterleitungsfunktionalität – so genannte Router. Router erledigen diese Weiterleitung von Datenpaketen durch das Netz, indem sie ein Datenpaket jeweils in die optimale Richtung schicken, damit das Datenpaket dem Ziel »näher kommt«. Wenn die kürzeste Wegstrecke jedoch verstopft ist (was leider sehr oft der Fall ist), kann die optimale Wegstrecke ein Datenpaket physikalisch vom Ziel entfernen. »Näher kommen« ist also in Bezug auf die Ankunftszeit zu sehen, und irgendwann kommen die meisten Daten beim Adressaten dann doch an.

Man sollte beachten, dass diese Weiterleitung für jedes einzelne Datenpaket individuell gemanagt wird. Da eine Webseite bei einer Übertragung in eine Vielzahl von einzelnen Datenpaketen zerlegt wird, kann das dazu führen, dass der eine Teil der Webseite einen anderen Weg nimmt als der andere. Im Extremfall nimmt der obere Teil der Webseite den Weg links um die Erde, während der untere Teil rechts herum zum Empfänger kommt. Dabei kann es auch vorkommen, dass zuerst in einer Webseite benötigte Datenpakete länger brauchen, um zum Empfänger zu gelangen, als darauf aufbauende. Etwa wie in der nachfolgend skizzierten Situation: Eine Webseite wird in hundert Datenpakete zerlegt, und die ersten fünfzig kommen in der richtigen Reihenfolge beim Empfänger an. Dann kommen alle fehlenden Pakete ab Nummer 52 an.



Nur Paket Nummer 51 wurde von einem Router auf einen Weg geschickt, der es nicht rechtzeitig ankommen lässt. Nummer 51 sieht sich vielleicht gerade noch die Welt an, während seine folgsamen Kollegen beim Browser vor der Tür stehen und sich langweilen. Der Browser wird so reagieren, dass er alle ohne Unterbrechung in der richtigen Reihenfolge auszuwertenden Datenpakete bereits verarbeitet und die Webseite so weit aufbaut, wie diese Datenpakete es erlauben. Ab der ersten Lücke lässt er dann die schon angekommen Datenpakete in Ruhe. Erst wenn der Urlauber da ist und sich in die Lücke einfügt, werden auch die bereits vor Langeweile ganz unleidlichen restlichen Datenpakete verarbeitet und die Webseite zu Ende aufgebaut. Sie erkennen solche Fälle daran, dass eine Webseite in einem Browser teilweise angezeigt wird, dann der Seitenaufbau lange Zeit nicht fortschreitet und plötzlich mit einem Schlag die Seite vervollständigt wird.

3.2.2 IP-Nummern und DNS-Namen

Jeder Rechner, der im Internet angeschlossen werden soll, bekommt eine eindeutige, numerische Adresse. Diese ist genau vier Byte lang und wird als IP-Nummer bezeichnet (IP steht wieder für Internet Protocol). Da in einem Byte 256 verschiedene Zeichen binär dargestellt werden können ($2 \text{ hoch } 8 = 256$), lässt sich in jedem der vier Byte eine Zahl zwischen 0 und 255 verschlüsseln. Gewöhnlich werden die IP-Nummern nicht binär, sondern im Dezimalsystem dargestellt und jedes Byte mit einem Punkt abgetrennt (zur besseren Lesbarkeit). Eine fiktive IP-Nummer wäre also so darstellbar: 123.132.133.178

Bei allen Datenübermittlungen im Internet (und auch im Intranet – also einem nicht öffentlichen Netzwerk, das Internettechnologie nutzt) werden solche IP-Nummern verwendet. Dennoch werden Anwender sie selten sehen. IP-Nummern sind zwar hervorragend dazu geeignet, in kompakter und eindeutiger Form eine Adressierung von Rechnern vorzunehmen. Für (normale ;-)) Menschen sind diese numerischen Adressen jedoch nicht einfach zu merken, zu verstehen und einzugeben. Glücklicherweise gibt es für viele der numerischen Adressen von Rechnern im Internet zusätzlich einen eindeutigen Alias-Namen, der alternativ verwendet werden kann.

Diese alternativen Adressnamen bilden ein System von logisch und inhaltlich zusammengehörigen Rechnergruppen, das hierarchisch geordnet ist. Logisch zusammengehörige Bereiche werden Domain genannt, woraus sich der Name für dieses Namenssystem ableiten lässt: DNS. Das ist nicht die Abkürzung aus der Biologie für Desoxyribonucleinsäure, sondern die Abkürzung für »Domain Name Service«, »Domain Name Server« oder »Domain Name System« (da gibt es drei offizielle Varianten – wichtig sind aber die ersten beiden Worte). Die einzelnen Bestandteile eines DNS-Namens werden wie IP-Nummern mit Punkten getrennt, die hierarchische Wertigkeit der Stufen ist jedoch von hinten nach vorne zu lesen. Der hinterste Teil eines solchen Alias-Namens stellt die größte logische beziehungsweise inhaltliche Einteilung da (die so genannte Top-Level Domain), etwa eine Nation oder eine Organisationsform. Für Rechner, die in Deutschland für das Internet

registriert sind², werden Sie dort beispielsweise DE finden, für Österreich AT und für die Schweiz CH. Der – unter Umständen mehrteilige – Mittelteil (der Second-Level) ist eine genaue Beschreibung des Rechnerverbandes bzw. des Rechners selbst. Er wird von dem Eigner einer Domain in der Regel möglichst sprechend gewählt. Etwa ZDF, ARD oder FFH, um populäre Medien als Beispiele anzugeben. Der vorderste Teil des DNS-Namens bezeichnet direkt den einzelnen Server auf einem Rechner bzw. Rechnerverband. Meist wird dafür WWW genommen, aber das ist in keiner Weise zwingend (nur sinnvoll, weil viele Internetanwender das auf gut Glück ausprobieren, und damit hat man weniger Besucher, wenn man davon abweicht). Gültige Alias-Namen in diesem DNS-Konzept sind folgende:

www.rjs.de

webscripting.de

www.mut.de

Vergeben werden diese Namen von der oben erwähnten Organisation NIC. Dabei kann jedermann oder jede Organisation einen solchen DNS-Namen eintragen, sofern dieser noch nicht von jemand anderem reserviert wurde. Diese Tatsache ist zwar bequem und einfach, birgt aber gewaltiges Konfliktpotential. Denn sehr oft kommt es zu Überschneidungen von Interessen. Was ist etwa, wenn zwei Städte mit gleichem Namen einen DNS-Namen für ihre Homepage reservieren wollen? Oder zwei Firmen mit gleichem Namen? Oder gar Personen, Vereine usw.? Bedauerlicherweise haben in der Vergangenheit diverse Geschäftemacher frühzeitig »gute« DNS-Namen (mit den Namen von großen Firmen als Second-Level) reserviert und nur für viel Geld an die Firmen abgetreten, als diese ihre Homepage bei NIC unter ihrem Firmennamen registrieren lassen wollten. Mittlerweile gehen die Geschäfte dieser »Domain-Grabber« allerdings schlecht, denn meist klagen Firmen oder Interessengruppen gegen – ihrer Ansicht nach – zu Unrecht belegte DNS-Namen und bekommen immer öfter Recht.³

Es gibt aber auch DNS-Namen, die als letzten Eintrag keine Bezeichnung der Nation haben. Das waren bis vor kurzem nur Rechner, die in den USA registriert waren. Der Grund ist einfach. Ursprünglich war das Internet ja auf die USA beschränkt, und damit machte eine Einteilung in Nationenkennungen keinen Sinn. Stattdessen wurde ein Regelwerk ver-

-
- 2 Die Registrierung eines Rechners hat überhaupt nichts damit zu tun, wo sich ein Rechner physikalisch befindet. So kann ein in Deutschland stehender Rechner auch in den USA registriert werden. DNS-Namen entsprechen in diesem Sinn Nummernschildern von Autos. Auch diese müssen selbstverständlich nicht dort bewegt werden, wo sie zugelassen sind, sondern können beliebig auf den Straßen umherfahren.
 - 3 So positiv diese Aktionen gegen Domain-Grabber auch sind – leider gibt es auch unzählige abstruse Urteile deutscher Gerichte, wonach DNS-Namen abgetreten werden müssen, obwohl dahinter echte Webseiten liegen und die Betreiber keinerlei Missbrauch beabsichtigt haben. Meist sind sie nur einer großen Firma oder einem Prominenten in die Quere gekommen, und dann gilt halt Macht vor Gerechtigkeit ;-(. Dies ist eines der vielen Zeichen, dass die deutsche Justiz mit dem System des Internets noch nicht klarkommt und damit fatalerweise keine Rechtssicherheit entstehen lässt.

wendet, bei dem im Wesentlichen sechs logische Bereiche benutzt werden.⁴ In der folgenden Tabelle sind sie alphabetisch aufgelistet.

Top-Level	Abkürzung für	Bedeutung
Com	commercial	Alle von kommerziellen Unternehmen betriebene Rechner
edu	educational	Rechner von Ausbildungsinstitutionen (Universitäten, Schulen etc.)
gov	government	Regierungsbehörden
mil	military	Militärische Computer
org	organization	Rechner von nicht kommerziellen Organisationen
net	network	Organisationen, die ein eigenes Netzwerk betreiben

Tabelle 3.1: Top-Level in den USA

Beispiele für gültige Namen nach diesem System sind folgende:

java.sun.com

www.opera.com

www.whitehouse.gov

www.unesco.org

Da in der letzten Zeit viele DNS-Namen bereits verwendet wurden, wurde Ende des Jahres 2000 eine Erweiterung der Top-Level Domains beschlossen, um Raum für neue DNS-Namen zu schaffen. Neu hinzugenommen wurden `info` (einzige offene neue Erweiterung; Markennamen werden bevorzugt), `name` (Domain-Erweiterung für Namen von privaten Personen; nur in Verbindung von Vor- und Nachname), `pro` (Freiberufler), `biz` (kommerzielle Unternehmen; Alternative zu `com`), `museum` (Museen), `aero` (Unternehmen der Luftfahrt) und `coop` (genossenschaftliche Unternehmen). Die technische Umsetzung des neuen Top-Levels ist aber zur Zeit der Bucherstellung⁵ noch nicht weit gediehen.

3.2.3 Uniform Resource Locator

Computer sind über IP-Nummern beziehungsweise ihre Alias-Namen eindeutig adressiert. Auf einem Computer und erst recht innerhalb des Internets müssen aber auch alle

4 Da dieses System ausschließlich in den USA verwendet wird, ist die Nation implizit klar. Beachten Sie aber wieder, dass auch jeder außerhalb der USA befindliche Rechner dort für »das Internet zugelassen« (d.h. registriert) werden kann und damit auch eine entsprechende Top-Level-Kennung erhalten kann.

5 Erste Hälfte 2001.

Daten und Programme in Form von eindeutigen Angaben adressiert werden. Damit werden neben der Rechneradresse zusätzlich einige andere Angaben notwendig. Im Wesentlichen Pfadangaben (also Verzeichnisse) und Dateinamen.

Der Name dieser vollständigen Adressangaben für Dokumente (im weitesten Sinn) im Internet lautet URL und steht für Uniform Resource Locator, was ins Deutsche übersetzt ungefähr einheitliches Adressierungsschema für Objekte im Internet bedeutet. Einheitlich deshalb, weil mit einer URL sowohl verschiedene Dienste, wie WWW, FTP, Gopher usw., als auch Rechner selbst oder direkt Dokumente (selbst auf dem eigenen Rechner – sogar ohne irgendwelche Netzwerkverbindung – über das Protokoll FILE) beschrieben werden können. Der Begriff Objekt steht in diesem Zusammenhang für so ziemlich alles, was Sie im Netz oder im eigenen Rechner finden. Die URL ist letztendlich das, was Sie in der Adresszeile des Browsers eintragen, um eine bestimmte Webseite angezeigt zu bekommen. Er ist aber auch die Angabe einer Datei, die Sie downloaden wollen, die lokale Datei, die im Browser angezeigt werden soll, oder auch der Aufruf des E-Mail-Programms mit einer vorbelegten E-Mail-Adresse. Die exakte Schreibweise einer URL ist je nach Dienstprotokoll leicht unterschiedlich, sieht jedoch in der Regel folgendermaßen aus:

```
<Dienstprotokoll>://<host:port></pfad></datei>
```

- Die Angabe Dienstprotokoll steht beispielsweise für http, ftp, gopher, file, news oder mailto. Danach folgen ein Doppelpunkt und fast immer zwei Schrägstriche zur Abtrennung (Ausnahmen sind mailto oder news).
- Mit host wird ein Server im Internet adressiert. Dabei kann direkt die IP-Nummer oder der DNS-Name eingegeben werden.
- Die Angabe port ist als Kanal zu verstehen. Auf einem Internetrechner bzw. Rechnersystem kann eine ganze Reihe von verschiedenen Diensten parallel betrieben werden. Das bedeutet, auf einem Rechner laufen mehrere Serverprogramme⁶ (üblich ist z.B. ein FTP-Server und ein HTTP-Server). Um auf einem ausgewählten Rechner den gewünschten Dienst zu erreichen, benötigt man also noch eine weitere Information. Dafür gibt es den so genannten Port bzw. Kanal. Im Rahmen von TCP sind 65.536 numerische Werte für einen Port möglich. Die Port-Adressen 0 bis 1023 sind dabei reserviert. Sie sind vordefinierten Diensten zugewiesen (»well-known ports«). Port-Adressen über 1023 sind aber durchaus möglich. Welcher Dienst darüber angesprochen werden soll, liegt in der Verantwortung des Anbieters. Das bedeutet, nahezu jeder Internetdienst hat einen Defaultwert, der immer dann verwendet wird, wenn man in einer URL keinen Port notiert. In der Regel braucht deshalb der Port nicht explizit angegeben zu werden. Der Default-Port für zum Beispiel einen HTTP-Server ist 80, ein FTP-Server hat den Port 21. Es ist allerdings möglich, die Ports für bestimmte

⁶ Beachten Sie, dass ein Server als Programm zu verstehen ist. Die oft benutzte Gleichsetzung mit Hardware ist irreführend und sollte nur in Spezialfällen verwendet werden.

Zwecke zu manipulieren, d.h., es ist optional möglich, in der URL mit `:port` festzulegen, auf welchem Kanal dieser Server angesprochen werden soll.

- Das genaue Objekt, das mit einer DNS-Angabe referenziert werden soll, verbirgt sich hinter der Angabe `</pfad></datei>`. Das ist eine übliche Pfad- und Dateiangabe im Rahmen der Verzeichnisstrukturen eines Rechners. Beachten Sie, dass diese gesamten Angaben von Unix abstammen und deshalb als Trennzeichen kein Backslash (wie unter DOS oder Windows), sondern ein Slash (also ein Schrägstrich in die andere Richtung) verwendet wird.

3.3 HTML und weitere Webtechniken

Wir beschäftigen uns in diesem Buch zwar mit dem Internet, nicht aber damit, wie Webseiten erstellt werden oder wie man im Internet programmiert. Dazu gibt es genügend andere Literatur. Nichtsdestotrotz ist ein kurzer Exkurs zu den Hintergründen der im Internet und vor allem im WWW verwendeten Techniken bzw. Technologien auch für reine Anwender sinnvoll. Da der Opera-Browser sehr flexibel konfigurierbar ist und sich die meisten WWW-Techniken einzeln aktivieren oder deaktivieren lassen, ist eine gewisse Grundlage an Detailwissen sinnvoll. Das Thema ist so umfangreich, dass wir auf die einzelnen Techniken genauer bei der Konfiguration von Opera eingehen. Eine Ausnahme ist HTML, die mit Abstand wichtigste Technologie im Web. HTML ist die Grundlage aller anderen im WWW verwendeten Techniken und lässt sich auch in einem Browser nie deaktivieren.⁷ Sie ist das Rückgrat aller Vorgänge in einer Webseite. Ohne HTML geht nichts im Web. Alle (!) anderen Techniken im WWW verankern sich letztlich irgendwo in einem HTML-Gerüst.

Wenn man das System von HTML einigermaßen verstanden hat, wird klar, warum viele Dinge im WWW so sind, wie sie sind. Und vor allem, warum sich viele Dinge in verschiedenen Browsern so unterschiedlich verhalten und ein und dieselbe Webseite so unterschiedlich aussehen kann.

HTML wurde aus der in der ISO-Norm 8779:1986 festgelegten Sprache SGML (Standard Generalized Markup Language) entwickelt und ist eine so genannte Dokument-Beschreibungssprache, mit der die logischen Strukturen eines Dokuments in plattformunabhängigem Klartext beschrieben werden. Sie ist keine (!) Programmiersprache wie Java, PASCAL oder C/C++, und mit HTML wird auch nicht programmiert (obwohl es häufig so genannt wird). HTML fehlen sowohl Variablen als auch Programmfluss-Anweisungen, die zentrale Bestandteile einer Programmiersprache sind. Im Grunde gibt ein Dokumentenformat nur unverbindliche Empfehlungen an eine interpretierende Darstellungssoftware (einen Browser wie Opera), wie eine bestimmte Dokumentenstruktur dar-

⁷ In einem E-Mail-Client jedoch schon.

zustellen ist, damit sie dem geplanten Layout und der vorgesehenen Funktionalität so weit wie möglich entspricht. Es gibt in reinem HTML keine absolute Darstellungsvorschrift, weswegen sich Darstellungen von HTML-Seiten in verschiedenen Browsern oft erheblich unterscheiden können. Das erscheint layoutfixierten Leuten immer als eine Schwäche von HTML, ist aber genau das Gegenteil. HTML erlaubt die plattformunabhängige Beschreibung einer grafischen Dokumentenseite mit einem äußerst geringen Aufwand und einer extrem kleinen Menge an zu übertragenden Daten. Nur darüber kann in einem – zeitweise unerträglich – langsamen Netzwerk wie dem Internet Information in einer adäquaten Zeitspanne übermittelt werden.

Zu einer durch HTML beschriebenen logischen Dokumentenstruktur gehören Verweise, aber auch Kapitel, Unterkapitel, Absätze usw. Da ein in HTML geschriebenes Dokument von Anfang an mit den Texten Grafiken sowie einige weitere multimediale Elemente (Sound, Video usw.) verknüpfen sollte, mussten entsprechende Techniken zur Verbindung mit einer Webseite entwickelt werden. Elemente, die jenseits von Textanweisungen zu sehen sind, werden in einer Webseite als Referenz auf eine entsprechende externe Datei notiert (eine solche Referenzierung kann in Opera sehr differenziert deaktiviert werden, was bei überladenen Webseiten den Ladevorgang erheblich beschleunigen kann). Wenn diese in einer Webseite dargestellt werden soll, muss natürlich die Präsentations-Software entsprechende Softwaremodule und die Hardware die zugehörigen Komponenten (beispielsweise eine Soundkarte für akustische Daten) verfügbar haben.

Eine weitere wichtige Eigenschaft von HTML ist, Verbindungen zu anderen Internetdiensten in eine Webseite aufnehmen zu können. Diese werden als Referenz in einer Webseite notiert, weshalb unter der Oberfläche des WWW viele Dienste wie E-Mail oder FTP verfügbar gemacht werden können. Es ist sogar eingeschränkt möglich, mit HTML Datenbankabfragen zu formulieren, die Resultate optisch aufzubereiten und Menüstrukturen aufzubauen. Auch Interaktion mit Anwendern ist – allerdings mit erheblichen Einschränkungen – in purem HTML zu realisieren.

HTML beinhaltet in der Version 1.0 wie gesagt nur wenige der heute im Web zu findenden Möglichkeiten. Da bereits der Mosaic-Browser diverse eigene Features implementiert hatte, die von dem einfachen HTML-1.0-Standard stark abwichen, und sich der Browser rasend schnell verbreitete, versuchte das W3C relativ zügig, einen um einige dieser neuen Möglichkeiten erweiterten HTML-Standard zu schaffen. Aber bereits damals zeigte sich, dass das W3C von den Browser-Herstellern nicht sonderlich ernst genommen wird. Der im November 1993 gestartete Versuch zur Einführung eines neuen HTML-Standards war nicht sehr erfolgreich, und es dauerte bis September 1995, bis endlich ein neuer Standard – die HTML-Version 2.0 – verabschiedet werden konnte.

Dieser 2.0-Standard enthielt zwar einige neue Gestaltungsmöglichkeiten für das WWW. Sie waren jedoch nur ein Teil der bereits möglichen HTML-Techniken, da sich durch die lange Pause zwischen den HTML-Versionen 1.0 und 2.0 verschiedenste browserspezi-

fische Dialekte gebildet hatten, unter denen eine Standardisierung des WWW auch heute noch leidet. Auch die nächste HTML-Version (3.0) war ein Flop. Sie wurde wegen der mangelnden Konsensfähigkeit der am W3C beteiligten Organisationen und Unternehmen nie offiziell verabschiedet. Erst Ende 1996 gab es einen neuen offiziellen Standard – die Version 3.2. Der HTML-Standard 3.2 hatte ungefähr ein Jahr Bestand. Ende 1997 fand die offizielle Verabschiedung eines neuen HTML-4.0-Standards statt, der auch heute noch aktuell ist (mit nur marginalen Erweiterungen). Im Grunde gilt bei allen offiziellen HTML-Standards, dass fast alle Browser-Hersteller immer nur das herausnehmen, was ihnen passt, und das weglassen, was ihnen nicht bequem erscheint. Oft scheint es so, als ob der eine Browser-Hersteller bestimmte Dinge nur deshalb auf eine bestimmte Art macht, damit sein Standard sich im Web etabliert und Besucher mit anderen Browsern möglichst viele Probleme bekommen.⁸ Das gilt leider nicht nur für HTML, sondern auch für andere Techniken wie JavaScript oder Style Sheets. Eine löbliche Ausnahme ist das Objekt unserer Begierde (in diesem Buch) – Opera. Der Opera-Browser hält sich an die offiziellen Vorgaben vom W3C!

3.3.1 Ein paar HTML-Grundlagen

Obwohl wir hier nicht die Erstellung von Webseiten im Blick haben, soll ein kurzer Einblick in die (!) Grundlagentechnik des Web vermittelt werden: HTML.

HTML ist als Klartext-Dokumentbeschreibungssprache viel einfacher als Programmiersprachen und auch Scriptsprachen. Es gibt beispielsweise keinerlei Kontrollstrukturen in Form von Bedingungen, Sprüngen oder Schleifen. Es gibt auch keine Befehle im Sinne von Befehlswörtern, die eine Aktion auslösen. HTML beinhaltet in seiner aktuellen Version nur einige Erweiterungen für den Aufruf von Scripten.

Da HTML eine äußerst fehlertolerante Beschreibungssprache ist, werden Webseiten selbst in Situationen, in denen in anderen Sprachen geschriebene Dokumente oder Programmstrukturen einen Fehler oder einen Programmabbruch auslösen würden, oft noch brauchbare Resultate liefern. Reine HTML-Seiten haben (bis auf Sonderfälle, die wir hier ignorieren) die Dateiendung HTM, HTML oder SHTML. Alleine diese lassen das Dokument zu einem HTML-Dokument werden. Für eine HTML-Seite gibt es jedoch wie für jedes andere Dokument oder jede andere Programmiersprache immer ein Grundgerüst und gewisse Grundregeln. Dabei sind besonders zwei Begriffe wichtig: Steueranweisungen und Tags.

HTML-Befehle werden gewöhnlich Steueranweisungen genannt. Diese Steueranweisungen werden in Form einer Stapeldatei in Klartext erstellt. Sprünge und Schleifen – im

⁸ Zweck: Der Anwender des anderen Browsers soll glauben, sein Programm wäre schlecht, damit er den Browser wechselt. Aber wie gesagt – vielleicht scheint es ja nur so ;-).

Source – kommen nicht vor, dafür sind die Layout-Möglichkeiten in HTML sehr ausgeprägt.

Alle HTML-Steueranweisungen werden in so genannte »Tags« geschrieben, die von spitzen Klammern – dem »kleiner«- und dem »größer«-Zeichen – begrenzt werden. Ein HTML-Tag sieht also von der Struktur her immer so aus: `<xyz>`

Dabei unterscheidet man zwischen Einleitungs- und Abschluss-Tags. Abschluss-Tags sind bis auf einen der öffnenden Klammer folgenden Slash (/) identisch zum Einleitungs-Tag. Der Abschluss-Tag zum obigen Einleitungs-Tag würde wie folgt aussehen: `</xyz>`

Beispiele:

```
<i> ... </i>
```

```
<h6> ... </h6>
```

```
<center> ... </center>
```

```
<p> ... </p>
```

Bei den HTML-Steueranweisungen spielt es im Gegensatz zu den Befehlsanweisungen vieler anderer Sprachen keine Rolle, ob sie groß- oder kleingeschrieben werden. Die Anweisung `<h4>` bewirkt das Gleiche wie `<H4>`. Auch eine unterschiedliche Groß- und Kleinschreibung im Einleitungs- und Abschluss-Tag oder sogar innerhalb einer Anweisung hat keine negativen Auswirkungen, erhöht jedoch nicht gerade die Lesbarkeit. Für die Zukunft steht aber zu erwarten, dass auch in HTML Groß- und Kleinschreibung relevant wird. Dann werden HTML-Anweisungen wahrscheinlich kleingeschrieben.

Es gibt Tags, die zwingend einen Einleitungs- und einen Abschluss-Tag benötigen. Beide zusammen bilden einen so genannten Container, wie in den obigen Beispielen. Andere Tags kommen in der strengen HTML-Syntax hingegen nur als Einleitungs-Tag vor. Beispiel (ein Zeilenumbruch): `
`

Das Grundgerüst einer HTML-Seite

Eine HTML-Seite wird immer in die Anweisung `<HTML>` am Anfang und `</HTML>` am Ende eingeschlossen. Die beiden Anweisungen bilden immer das äußere Gerüst einer HTML-Seite. Davor dürfen höchstens Kommentarzeilen stehen, die natürlich auch überall im Inneren einer HTML-Seite verwendet werden können. Die Steuerzeichen für Kommentarzeilen sind ein Ausrufezeichen und zwei Striche am Anfang und zwei Striche am Ende des Tags. Jeder in diesem Tag stehende Text wird vom interpretierenden Browser als Kommentar betrachtet. Ein Kommentar kann über mehrere Zeilen gehen und natürlich ebenfalls im Inneren einer HTML-Seite verwendet werden. Auch ein Kommentar-Tag muss von spitzen Klammern eingeschlossen werden, also sehen die Zeichenfolgen so aus: `<!-- ... -->`

Beispiel: `<!-- Generated by RJS - Last Updated: May 28, 2001 -->`

Das weitere Grundgerüst einer HTML-Datei besteht grundsätzlich aus folgenden zwei Teilen:

- Dem (optionalen) Header (Kopf)
- Dem Body (Körper)

Nach der einleitenden HTML-Anweisung kommt auf einer Webseite normalerweise ein Header-Teil, d.h. ein Kopfteil, in dem die allgemeinen Informationen über die Seite zusammengefasst werden. Beispielsweise der Titel, welcher in die Zeichenketten `<TITLE>` und `</TITLE>` eingeschlossen wird. In diesem Bereich stehen etwa Copyright-Informationen und vor allem wichtige Angaben für Suchmaschinen. Ein Header wird mit dem Steuerzeichen `<HEAD>` begonnen und mit `</HEAD>` entsprechend wieder geschlossen.

Die eigentlichen Daten, die der WWW-Browser einem Anwender auf dem Bildschirm anzeigen soll, werden in den Body geschrieben. Die Tags `<BODY>` und `</BODY>` umschließen den Anzeigebereich einer Webseite. In dem Body ist der eigentliche Text mit Überschriften, Verweisen, Grafikreferenzen und auch die Referenz für Java-Applets zu notieren.

Das vollständige Grundgerüst einer normalen HTML-Datei sieht also schematisch immer so aus.

```
<!-- Kommentar -->
<HTML>
  <HEAD>
    <TITLE>Titel</TITLE>
  </HEAD>
  <BODY>
    Überschriften, Text,
    Verweise, Grafiken, Java-Applet-Referenz usw.
  </BODY>
</HTML>
```